MassTrack: The MBTA Green Line

TONY FALLER (AF3370), Columbia University, USA

1 SYNOPSIS

MassTrack is an analytical project aimed at evaluating the privacy implications of public transit data. Within the context of another course (COMS E6998: Private Systems), research on BlueBike transit data has demonstrated that combinations of particular fields, particularly those containing fine-granularity timestamps, serve as strong pseudo-identifiers with near 100% confidence (see Appendix B). This same analysis, when applied to Massachusetts Bay Transit Authority (MBTA) data, produces extremely dissimilar results.

MBTA transit data exhibits significant noise. Analysis of the combination (start station, start time, direction) reveals about 2% uniqueness across 2024, suggesting that individual journeys are far less distinguishable compared to BlueBikes.

To explore a possible attack vector, MassTrack introduces a model that reconstructs train rides based on available inputs: start station, start time, and time spent in transit. This system maps real-life MBTA data records and outputs sequential station/timestamp pairs, documenting a journey. However, because this combination is rarely unique, such reconstructions are not widely applicable, demonstrating that this dataset is vulnerable under only very specific conditions.

The comparative takeaway is clear: MBTA transit offers a higher degree of anonymity than BlueBikes, where timestamp precision makes re-identification easier. Through MassTrack's dual approach, this project provides insights into transit privacy while quantifying pseudo-identifier strength in different commuting systems. These results are valuable for both privacy researchers and privacy-conscious commuters.

MassTrack is available as a pair of open-source Jupyter notebooks. The MBTA analysis can be found at https://github.com/amfaller/MassTrack-MBTA and the BlueBike analysis can be found at https://github.com/amfaller/MassTrack-BlueBike.

1.1 Datasets & Benchmarks

The MBTA publishes transit data online (see https://www.mbta.com/developers). Due to computational power constraints, the scope of this analysis is limited to 2024 Green Line data. See Appendix A for further information regarding the Green Line.

Author's address: Tony Faller (af3370), af3370@columbia.edu, Columbia University, New York, USA.

This dataset reflects station-to-station transit data, with fields such as start station/time, end station/time, and route id. These fields are given in human-readable text and General Transit Feed Specification (GTFS) formats. Timestamps have second-level granularity.

This dataset, including its data dictionary, is publicly available here: https://mbta-massdot.opendata.arcgis.com/datasets/0b4dc16b8b984836962229865d5b573b/about. For the purposes of this analysis, this dataset was re-uploaded to Zenodo and is available here: https://zenodo.org/records/15121997. Due to time constraints and a lack of ground-truth data, there was no benchmark used against this dataset.

1.2 Research Questions

(1) *RQ1: Does the combination of (startStation, startTime) constitute a pseudo-identifier?*
(2) *RQ2: Does the combination of (startStation, startTime, direction) constitute a pseudo-identifier?*
(3) *RQ3: How accurate is the MassTrack system?*
(4) *RQ4: How does time-in-transit affect the execution time of the MassTrack system?*
(5) *RQ5: How does time-in-transit affect the accuracy of the MassTrack system?*

2 IDENTIFYING PSEUDO-IDENTIFIERS

Pseudo-identifiers are datapoints / fields which, when combined, can uniquely re-identify a row in a dataset. If this dataset is composed such that one row corresponds to one person, this is functionally equivalent to re-identifying this individual. In the BlueBike study, timestamp precision was shown to strongly affect re-identification risk; the same analysis is applied to MBTA data via RQ1 and RQ2.

2.1 Methodology

The methodology for both of these questions is as follows:

(1) Use *Pandas df.groupBy()* to aggregate records based on the selected attributes
(2) Identify groups containing exactly one record – these correspond to uniquely-identifiable trips
(3) Calculate the percentage of total groups which have exactly one record:

$$\text{Uniqueness Percentage} = \left(\frac{\text{Number of groups with 1 record}}{\text{Total number of groups}}\right) \times 100$$

A higher uniqueness percentage suggests that this combination of fields is a strong pseudo-identifier; conversely, a lower uniqueness percentage suggests that this combination of fields is a weak pseudo-identifier.

2.2 Results & Findings

Across 2024, uniqueness of these fields was found to only be about 1-2%. Direction did not significantly impact uniqueness; plots differ only slightly in decimal values.

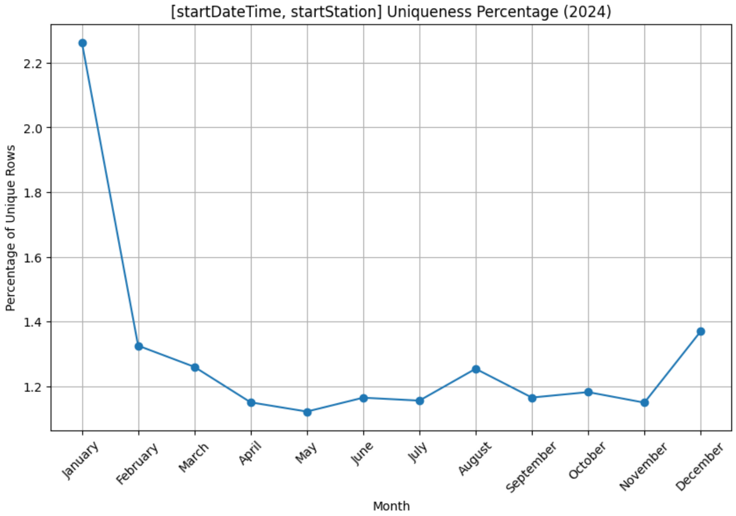

Fig. 1. Uniqueness of records given start station & start time.

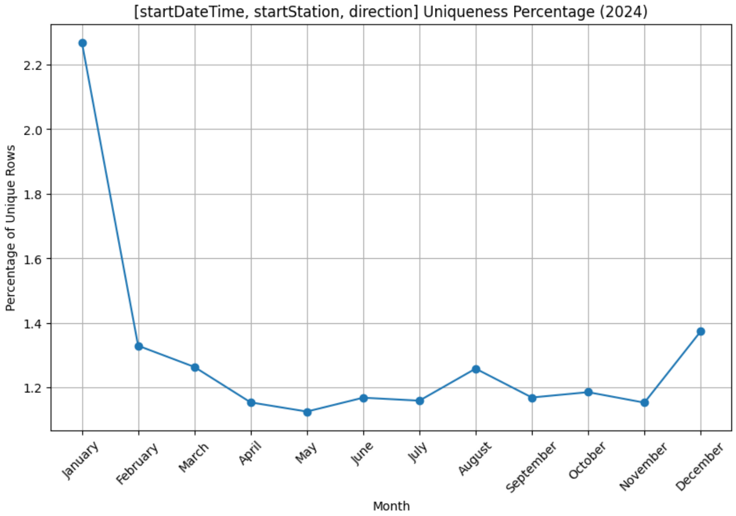

Fig. 2. Uniqueness of records given start station, start time, & direction.

3 TRACING A RIDE

The MassTrack system attempts to reconstruct a train journey given inputs of start station, start time, and time spent in transit, leveraging real-world MBTA data to list every station visited during the journey. If this system has a high degree of accuracy (RQ3), it demonstrates that an adversary may leverage this dataset for the purposes of tracking, which is a privacy invasion.

Given a lack of ground-truth data to compare against, assessing accuracy becomes difficult. However, this can still be evaluated through logical station ordering (i.e. does the sequence of stations in the journey make sense?), realistic timing (does the journey exhaust the time-in-transit parameter?), and consistency in the produced journey across iterations of fixed inputs.

3.1 Methodology

3.1.1 Journey Reconstruction. The MassTrack system accepts inputs of start station, start time, and time spent in transit. MassTrack instantiates a "transit budget" equivalent to the time-in-transit parameter; then:

(1) Searches for a row in the dataset which matches the input start time and station
(2) Searches for the next row which:
    (a) Has a later timestamp (if multiple, choose the closest to the previous timestamp); and
    (b) Is going in the same direction; and
    (c) Is on the same sub-Green Line as the previous row
(3) Subtract the difference between timestamps from the transit budget
(4) Repeat until budget is exhausted or no next ride can be found

3.1.2 Test Case Generation. Five random test cases are generated by selecting a random row from the dataset to collect a real start station and start timestamp. Then, a random time-in-transit between 25 and 90 minutes is assigned. These values are input to the MassTrack system.

The specific test cases used to evaluate the MassTrack system are available within the linked GitHub repository.

3.2 Evaluation Criteria, Results & Findings

3.2.1 Logical Ordering. Verify that the sequence of stations output by the MassTrack system follows the expected station order within the MBTA transit network. Report the number of missing interim stations across journeys.

For five randomly-generated test cases, all stations were sequential, but journeys contained gaps; there were an average of 12.2 missing stations per journey. This constitutes nearly half of the stations one would pass through on the longest possible Green Line trips. This is partially explainable with the notion of express trains – when one train catches up with another which is ahead of it, the farther-along train will skip some future stations to reset inter-train spacing.

There is no way to know ahead of time if this happens, and this scenario is not explicitly marked within the dataset. Anecdotally, it is unlikely for an express train to skip more than two or three stations at a time.

3.2.2 Realistic Timing. Report the amount of time-in-transit which is not utilized by the MassTrack system, i.e. is unaccounted for within the journey.

With an average starting budget of 49.6 minutes, there was an average of 17.5 minutes which MassTrack did not consume. This is an average of 25% of the starting budget, but this may not reflect real-world journeys as these time-in-transit values were random. This finding would imply that most trips take an average of 32.1 minutes, though it is possible that a larger time budget is required to capture a full journey.

3.2.3 Consistency. Run a test case once, then ten more times. Compare these ten additional iterations' output journeys for consistency with the first execution. MassTrack demonstrated 100% consistency across all test cases.

4 MASSTRACK PERFORMANCE

Execution time directly affects MassTrack's usability for real-world analysis. Vulnerabilties may or may not exist, but existing vulnerabilities would be tempered with a long execution time – this would reduce the ease with which this dataset can be exploited. Because other parameters are limited to the options available within the dataset, the independent variable is the time-in-transit parameter (RQ4, RQ5).

To assess execution time and accuracy as a function of time-in-transit, test cases were generated as was done in the previous section.

4.1 Execution Time

Each test case was executed iteratively, with time-in-transit values ranging from 40 to 115 minutes in 5-minute increments. Execution time was measured across all executions and averaged to produce the following plot.

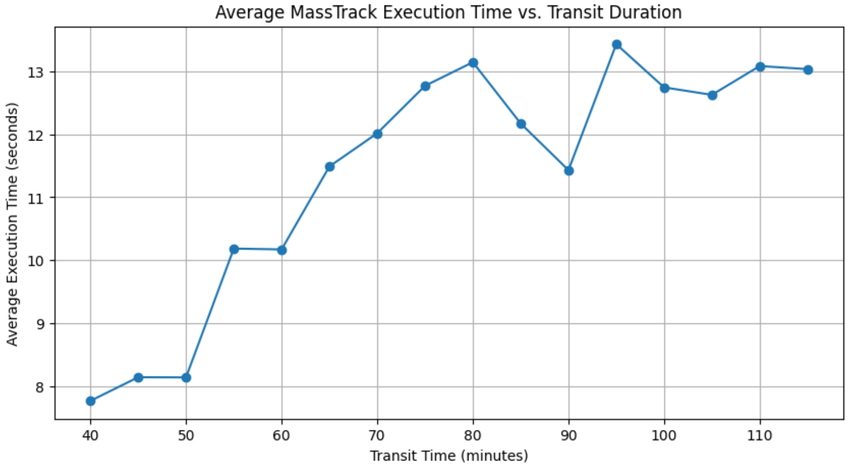

Fig. 3. Average execution time as a function of time-in-transit.

Execution duration appears to asymptotically approach 13 minutes after a time-in-transit value of 80 minutes. Prior to this threshold, execution time appears to generally increase.

4.2 Accuracy

As was done in RQ3, each test case was executed 11 times per time-in-transit value. Similar to RQ4, the time-in-transit value ranged from 40 to 115 minutes in 5-minute increments. It was found that across all test cases, MassTrack was 100% consistent across all time-in-transit values.

5 CONCLUSION

The goal of this project was to examine a publicly-available dataset and report privacy vulnerabilities as they relate to pseudo-identifiers. A system known as MassTrack was developed which leverages real-world train data to reconstruct real-world journeys. However, the quality of the analyzed dataset is poor; as a consequence, the data is safe from a privacy perspective, and MassTrack cannot easily be used for a re-identification attack.

Uniqueness is low because the dataset is noisy. It was found that MassTrack output journeys which had many missing interim stations. These interim stations did not have corresponding records, and the cause of this is unknown. Perhaps there are gaps in data recording, or perhaps this dataset was santized pre-release.

Some record groupings were found which indicate 30 or more trains leaving the same station in the same minute. In some cases, the real-life station is small and only contains enough physical space for one train in each direction.

A BlueBike rider is more vulnerable to a privacy invasion, as compared to a T rider. The combination of (start station, start time) is sufficient to re-identify a BlueBike rider with 100% confidence; however, the same fields can re-identify an MBTA rider with only 2% confidence. Additional knowledge of the direction in which the train is traveling does not have a significant impact.

5.1 Self-Evaluation (What I Learned)

The largest challenge was managing the MBTA dataset. This file is provided by the MBTA as a single .zip file containing 24 .zip files – two for each month, with each month corresponding to "light rail" and "heavy rail" data files. The size of the top-level .zip file is 1.3 GB; though I never processed the data locally, simply extracting all of this data is more than enough to overload the RAM of the free tier of Google Colab.

By comparing these results against the results of the BlueBike portion of the study, it is proven that renting a BlueBike exposes a rider to a much higher degree of privacy risk (100% reidentification risk, knowing the timestamp) as compared to taking the T (2% reidentification risk, knowing timestamp, station, and direction).

Future work should examine additional years to better understand if this safety is confined to the 2024 dataset. Additionally, a justification should be developed for the noise within the dataset. Finally, ground-truth data should be used to better assess the accuracy of the MassTrack system.

5.2 Who Did What

MassTrack (both its BlueBike and MBTA portions) was executed as a solo project. This project was split between two courses, but there was no overlap in instantiation or implementation. Results from each are included in the other for comparison.

A THE GREEN LINE

Boston's public transit system is operated by the Massachusetts Bay Transportation Authority (MBTA). The subway system is known as the "T" and is organized into color-coded lines. Origin-destination (O/D) data for these lines is publicly available. This project focuses on the Green Line, which itself splits into several letter-named branches:

Line	Route
B Line	Boston College ↔ Government Center
C Line	Cleveland Circle ↔ Government Center
D Line	Riverside ↔ Union Square
E Line	Heath Street ↔ Medford/Tufts

Table 1. MBTA Green Line Routes.

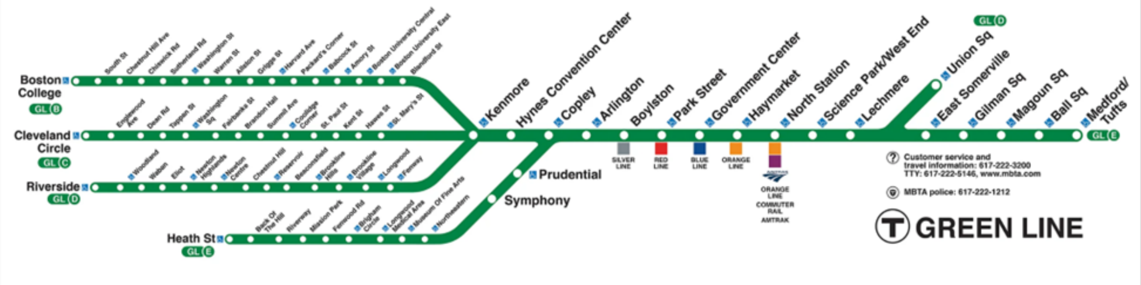

Fig. 4. Boston's Green Line, part of the T subway system.

Structural Observations

- **West of the Kenmore stop:** The B, C, and D lines are isolated.
- **East of the Kenmore stop:** The B, C, and D lines all run on the same track and stop at the same stations.
- **Copley Exception:** Both of the above are true for the E line, except at the Copley stop instead.
- **East of Government Center & West of Lechmere:** The D and E lines run on the same track.
- **East of Lechmere:** The D and E lines are isolated.

Abstracting the Green Line

A graph can be constructed to represent the Green Line. Each node is a station, and edges represent connections between stations. Edge weights represent the transit time between stations, based on MBTA data.

Transfer Nodes. Stations at which a rider is likely to transfer to a different train, rather than remaining on their current train, are designated as **Transfer Nodes**. This typically occurs when a rider needs to change directions or lines. For example, a passenger traveling from BU East on the B Line to Waban on the D Line would likely transfer at Kenmore.

Important Considerations

While inter-line transfers (to the Red, Orange, or Blue Lines) exist, this analysis assumes riders remain entirely within the Green Line.

B BLUEBIKE RESULTS

This appendix summarizes selected findings from the Private Systems portion of the MassTrack project. Further detail is available at: https://github.com/amfaller/MassTrack-BlueBike.

Key Takeaways

(1) Coarse timestamps, when combined with at least one other field (e.g. start station), can serve as a strong pseudo-identifier.

(2) Precise timestamps can function as strong pseudo-identifiers on their own.

BlueBike Research Question 1

Given start station/time, can end station/time be uniquely determined?

- **Known Information:** Start station, start time within N minutes.
- **Possible Gains:** End station, precise start time, precise end time.
- **Parameter(s):** N - Start time granularity.
- **Efficacy Metric:** Percentage of unique rows (higher percentage → stronger pseudo-identifier).

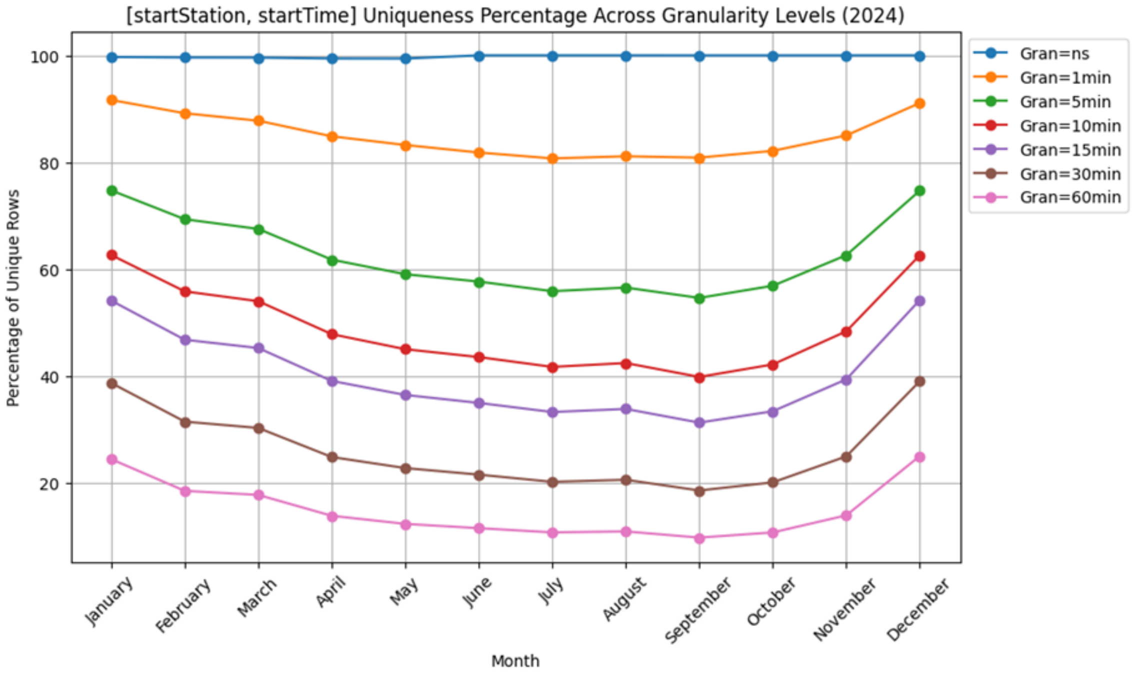

Fig. 5. Timestamp granularity was shown to greatly impact re-identification confidence.

BlueBike Research Question 7

Is (Start Time) a pseudo-identifier?

- **Known Information:** Start time within N minutes.
- **Possible Gains:** Start station, end station, precise start time, precise end time.
- **Parameter(s):** N - Start time granularity.
- **Efficacy Metric:** Percentage of unique rows (higher percentage → stronger pseudo-identifier).

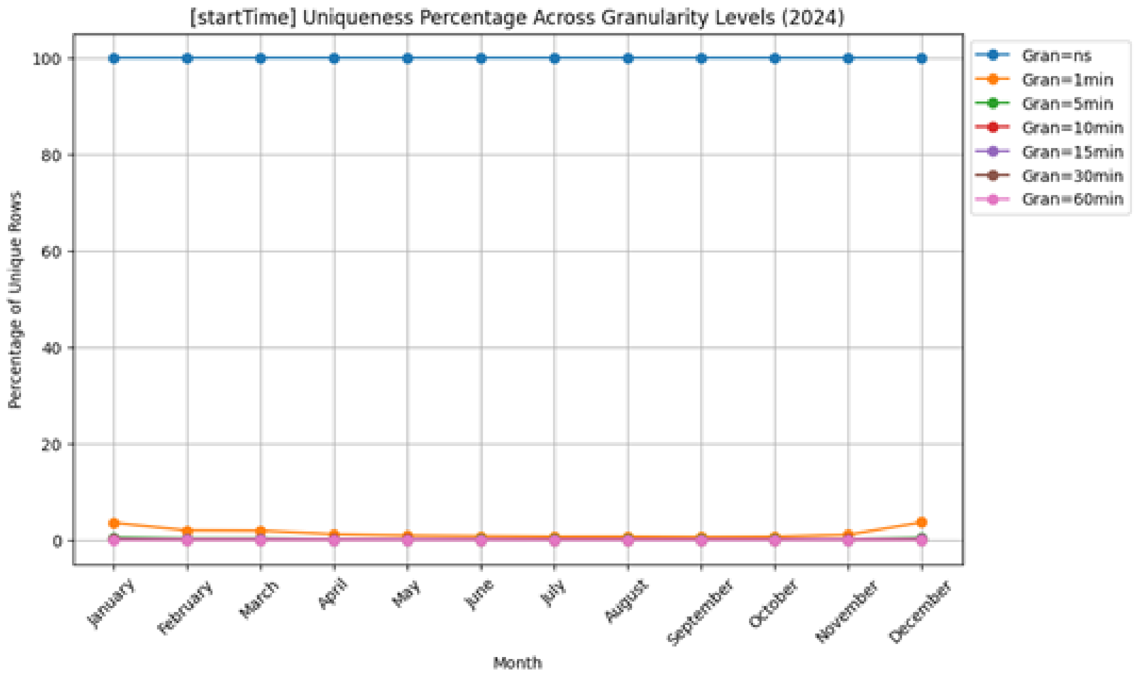

Fig. 6. Precise timestamps can act as pseudo-identifiers.